Too Smooth to Be Measured: What a Noise-Floor Provenance Test Says About a Published Spectral Library

Satyajeeth Suresh Kannan
Innovation Academy · Mary's Lab
github.com/JeetuSK0808/maryslab

Preprint · August 2026 · Mary's Lab technical report TR-13

ABSTRACT

A spectral dataset arrives labelled as measured reflectance. The label is a claim about provenance, and provenance is exactly what a table of numbers does not carry. The noise might: real measurements retain instrument noise that is small, roughly uncorrelated between adjacent wavelengths, and never absent, while modelled or heavily processed curves do not. Separating a spectrum into a smooth component and a residual with a Savitzky–Golay quadratic fit, we compute statistics that on constructed cases cleanly separate a noisy curve (0.811) from a smoothed one (0.667) and an analytic model (0.667). We then ran the test on 1752 reflectance spectra from the USGS Spectral Library Version 7 and it did not reproduce: the median published spectrum scores 0.678, statistically indistinguishable from the noiseless model, and the noise-floor flag fires on 1630 of them. We show this is not an artifact of our 10 nm grid — the same spectra score *lower* at native spectrometer resolution (0.667 against 0.673, up to 4,595 channels) — but a property of the library, which publishes processed data. The useful conclusion is a negative one: for a curated library, "these are measurements" is a claim about documentation that cannot be recovered from the numbers.

Keywords: data forensics; provenance; Savitzky–Golay; spectral datasets; research integrity

1. Introduction

Spectral libraries are cited as evidence. Results rest on the data being what its label says, and that assumption is almost never checked, because checking it has required a level of manual attention nobody has for thousands of records.

Files are resampled, smoothed for presentation, interpolated onto new grids, and occasionally generated from models. Each of those operations leaves the column headers unchanged. The question of whether a given curve was measured is therefore open by default.

We argue it is answerable, because measurement leaves a signature that processing removes.

2. Method

Any real measurement carries instrument noise: detector shot noise, electronics, calibration residue. Its magnitude varies with instrument and integration time but is never zero, and it is roughly uncorrelated between adjacent wavelengths. Modelled curves have no such component; heavily processed curves have far too little, and what remains is structured rather than random.

The difficulty is that the true underlying spectrum is unavailable, so the split into signal and noise must be made from the curve alone. A Savitzky–Golay filter fits a low-order polynomial over a sliding window and takes the fitted value at the centre. The choice matters: a quadratic fit over a short window reproduces the curvature of real absorption features rather than flattening them, so the residual retains noise without absorbing real structure. A moving average would place real features in the residual and make genuine spectra look noisier than they are.

From the residual and the curve we compute roughness, high-frequency energy ratio, second- and third-difference standard deviations, the ratio of distinct values (which detects quantisation and interpolation), and flat-tail length. These combine into a measured-likeness score in [0,1], reported alongside the individual metrics and any specific artifacts detected.

3. Evaluation

Inputs are synthetic with known provenance — necessarily, since evaluating a provenance detector requires knowing the provenance. All three curves derive from one analytic Gaussian reflectance, so shape is held constant and only provenance varies.

Table 1. Provenance statistics for three curves derived from one underlying spectrum. Only the processing history differs. The artifact column reports the flag raised and its severity in [0,1].

curve	measured-likeness	HF energy ratio	roughness	distinct ratio	artifact
measurement-like (noise added)	0.811	0.00173	0.02236	1	NO_NOISE_FLOOR@0.568
smoothed three times	0.667	0	0.00828	1	NO_NOISE_FLOOR@1
analytic model (no noise)	0.667	0	0.00867	0.553	NO_NOISE_FLOOR@1

The measured curve separates from both others, and the separation is driven by the high-frequency energy ratio exactly as the construction predicts: the noisy curve retains residual energy at 0.00173, while three passes of smoothing and the analytic model both report zero to the precision recorded. A curator screening a library can rank by this score and inspect the tail rather than inspecting everything.

What the score does not do is distinguish the two non-measurements. Smoothed and modelled both land at 0.667, and the noise-floor artifact saturates at severity 1 for both while reading NO_NOISE_FLOOR@0.568 for the measurement. This is the method working as designed rather than failing: the statistic is built on the absence of a noise floor, and both curves are missing it for different reasons that the statistic cannot see. Separating them requires a different signal, and the distinct-value ratio is a candidate — it reads 0.553 for the analytic curve against 1 for the smoothed one, because the analytic curve's tails repeat the same clamped value — but on one case we report that as an observation, not as a discriminator.

One further result deserves stating because it cuts against the instrument. The noise-floor artifact fires on the measurement-like curve too, at severity NO_NOISE_FLOOR@0.568. The synthetic noise amplitude used here sits below what the detector expects of a real instrument, so a genuine low-noise measurement would also be flagged. The graded severity carries the useful information; the presence of the flag alone does not.

What the numbers do not license is an accusation. Smooth is not fraudulent: publishing smoothed data is standard practice, and modelled spectra are useful and normal. What is wrong is smoothing or modelling and *labelling the result a raw measurement*. The audit detects the processing; whether the label is wrong is a question about the label, and the instrument reports findings rather than verdicts for precisely this reason.

Reproducing these numbers

Clone `github.com/JeetuSK0808/maryslab` and run `node papers/build/experiments.mjs`. Every figure in this report is written by that script into `papers/build/results.json`; the instrument itself is exercised by `audit_spectrum_provenance`. The engines carry a 401-vector conformance suite shared by a TypeScript reference implementation, a C++20 core, and a WebAssembly build, so a reported number is a property of the specification rather than of one implementation.

4. Real-data evaluation: the detector against a published library

The evaluation above is constructed, and a constructed evaluation cannot answer the question the method was built for: what does the detector say about spectra that a real library publishes as measurements? Snapshot `usgs-splib07a-1` makes that answerable — 1752 reflectance spectra from the USGS Spectral Library Version 7, public domain, gated Q1–Q6 on ingest. We scored every one.

The answer was not the one we expected.

Table 2. Measured-likeness over every record in the snapshot, against the two constructed reference curves from Table 1. Higher means more consistent with a raw measurement.

curve	n	mean	median	p10	p90
USGS splib07a, all records	1752	0.719	0.678	0.668	0.851
constructed “measurement-like”	1	0.811	0.811	—	—
analytic model, no noise	1	0.667	0.667	—	—

The median real measurement scores 0.678. The noiseless analytic curve scores 0.667. Those are the same number to within 0.011. Our own synthetic stand-in for a measurement scored 0.811 — higher than 1210 of 1752 real records. The `NO_NOISE_FLOOR` artifact fires on 1630 of them.

Read naively, the instrument has just accused the most widely cited spectral library in the field of not containing

measurements. That reading is wrong, and locating why is the contribution of this section.

Ruling out our own pipeline

The first suspect was us. The canonical grid is 38 samples at 10 nm; ASD spectrometers sample at roughly 1–2 nm, and channel-to-channel noise is not representable at 10 nm at all. If our resampling were erasing the noise floor, the finding would be an artifact of this project rather than a fact about the library.

That is a testable claim, so we tested it: score the same spectra at native instrument resolution, directly from the splib07a ASCII distribution, and compare against the same spectra resampled to the canonical grid.

Table 3. The same spectra scored at native spectrometer resolution and on the canonical grid. If the canonical grid were destroying the noise floor, the native column would score higher.

resolution	channels	mean	median	median HF ratio	n ≥ 0.7
native (splib07a ASCII)	224–4595	0.679	0.667	0	22
canonical grid	38	0.715	0.673	0.00008	105

The native column scores *lower*, not higher: median 0.667 against 0.673, on n = 400 records at up to 4,595 channels. Median high-frequency energy at native resolution is 0. Resampling, if anything, introduces a little high-frequency content rather than removing it. Our pipeline is not the cause, and the hypothesis that motivated this experiment is refuted by it.

What the finding actually is

The remaining explanation is the correct one, and it is documented in the library's own publication: splib07a spectra are published after processing. Records are spliced across spectrometers, corrected against reference standards, and resampled; several are explicitly convolved to other instruments' bandpasses. A published spectral library is a curated product, not a detector dump, and it carries the noise signature of one.

So the instrument is behaving correctly and the sentence it supports is narrower than the one a reader might reach for. It detects *processing*, and essentially every record in a major public library has been processed. It cannot distinguish a processed measurement from a model, because after enough processing there is no signature left to distinguish them by — the median native USGS spectrum and an analytic Gaussian are the same number to this instrument.

Three consequences follow, and they are more useful than the result we set out to get. First, the detector's operating point is wrong for library screening: a threshold that flags 1630 of 1752 records flags everything and ranks nothing. Second, it is nonetheless correctly ordered — the 1 record scoring below 0.5 and the tail at p10 = 0.668 are the ones worth a curator's attention, and relative ranking survives even when the absolute threshold does not. Third, and most importantly for anyone citing spectral data: the claim "these are measurements" is, for this library, a claim about *provenance documentation* and not one that can be recovered from the numbers. The numbers no longer contain it.

We note the other artifacts for completeness, since they fire rarely and therefore do carry information: `FLAT_TAIL` on 2 records, `CLIPPED_RANGE` on 2, and `QUANTIZATION_LATTICE` on 1. A flag that fires on three records out of 1752 is a flag worth reading.

Reproducing these numbers

Clone `github.com/JeetuSK0808/maryslab` and run `node papers/build/experiments.mjs`. Every figure in this report is written by that script into `papers/build/results.json`; the instrument itself is exercised by `audit_spectrum_provenance`, then `papers/build/provenance-resolution.mjs`. The engines carry a 401-vector conformance suite shared by a TypeScript reference implementation, a C++20 core, and a WebAssembly build, so a reported number is a property of the specification rather than of one implementation.

5. Limitations

This section states what the result does not establish. It is written to be read before the abstract is quoted.

- **Detects processing, not misconduct.** A low score indicates a curve does not carry measurement noise. The inference to mislabelling requires knowing what the label claimed, and the inference to intent is not available at all.
- **Cannot identify which operation was applied.** Low residual energy is consistent with a moving average, a Gaussian filter, a spline fit, or simulation from first principles. The audit narrows the space; it does not name the operation.
- **Passing is not correctness.** A well-calibrated instrument measuring the wrong sample produces con-

vincing noise on top of an incorrect curve. This tests one property.

- **The real-data section reports a failure of the operating point, not a validation.** On 1752 published spectra the score does not separate measurements from models, because the library ships processed data. §4 establishes that this is a property of the library and not of our resampling; it does not rescue the threshold. Anyone applying this instrument to a curated library should use it to rank, not to judge.
- **No library with documented raw provenance was available.** Showing that the detector works as designed would require spectra published straight from an instrument with the processing history recorded. We could not obtain one, so the premise — that raw measurements score high — remains supported only by synthetic noise we added ourselves, which is weak evidence and is labelled as such.
- **Instrument-dependent noise floors are not modelled.** The noise-floor threshold is a fixed expectation, and it fired on our measurement-like curve at partial severity. A genuinely low-noise instrument would produce curves that score as processed. Calibrating the threshold per instrument would require measured data from that instrument, which is the licensing constraint this project is under.
- **Prior art was surveyed informally.** Signal-quality metrics are standard; a search found no tool packaging them as a provenance screen for spectral libraries.

6. Acknowledgements and disclosure

The instruments described here were implemented in Mary's Lab, an open-source computational color science project. Software design, implementation, and the preparation of this report were carried out with substantial assistance from a large language model (Anthropic Claude), under the author's direction and review. All numerical results were produced by executing the released code; none were generated by a language model. The author is responsible for the claims made here.

References

1. Savitzky, A. and Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8), 1627–1639.
2. Farid, H. (2009). Image forgery detection. *IEEE Signal Processing Magazine*, 26(2), 16–25.
3. Wyszecki, G. and Stiles, W. S. (1982). *Color Science: Concepts and Methods, Quantitative Data and Formulae*, 2nd edition. Wiley.
4. CIE (2004). *Colorimetry, 3rd edition*. CIE 15:2004. Commission Internationale de l'Éclairage, Vienna.
5. Wilkinson, M. D., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018.
6. Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, 533, 452–454.
7. Schanda, J. (ed.) (2007). *Colorimetry: Understanding the CIE System*. Wiley-Interscience.

Source, engines, and the script that produced every number in this report: `github.com/JeetuSK0808/maryslab`. Results regenerate with `node papers/build/experiments.mjs`. Inputs: synthetic where §3 says so, and measured USGS splib07a spectra (snapshot `usgs-splib07a-1`) where a real-data section says so. Prepared with AI assistance (see Acknowledgements).